problems that lie ahead [101]. They identify several serious deficiencies of current systems. For example, termination criteria are still poorly understood. Although INTERNIST can diagnose simultaneous diseases, it also pursues all abnormal findings to completion, even though a clinician often ignores minor unexplained abnormalities if the rest of a patient's clinical status is well understood. In addition, although some of these programs now cleverly mimic some of the reasoning styles observed in experts [14],[48], it is less clear how to keep the systems from abandoning one hypothesis and turning to another one as soon as new information suggests another possibility. Programs that operate this way appear to digress from one topic to another — a characteristic that decidedly alienates a user regardless of the validity of the final diagnosis or advice.

9 Conclusions

This review has shown that there are two recurring issues to confront in considering the field of computer-based clinical decision making:

(1) How can we design systems that reach better, more reliable decisions in a broad range of applications, and

(2) How can we more effectively encourage the use of suth systems by physicians or other intended users?

We shall summarize by reviewing these points separately.

Performance Issues

Central to assuring a program's adequate performance is a matching of the most appropriate technique with the problem domain. We have seen that the structured logic of clinical algorithms can be effectively applied to triage functions and other primary care problems, but they would be less naturally matched with complex tasks such as the diagnosis and management of acute renal failure. Good statistical data may support an effective Bayesian program in settings where diagnostic categories are small in number, non-overlapping, and well-defined, but the lack of higher level domain knowledge limits the effectiveness of the Bayesian approach in more complex patient management or diagnostic environments. A mathematical approach may support decision making in certain well-described fields in which observations are typically quantified, and related by functional expressions. These examples, and others, demonstrate the the need for thoughtful consideration of the technique most appropriate for managing a clinical problem. In general the simplest effective methodology is to be preferred, but acceptability issues must also be considered as discussed below.

It is also always appropriate to ask whether computer-based approaches are needed at all for a given decision making task. The clinical algorithm developers, for example, have almost uniformly discarded the machine, and Schwartz et al. pointed out that a useful decision analysis can often be accomplished in a qualitative manner using paper and pencil [87].

Finally, it is important to consider the extent to which a program's "understanding" of its task domain will heighten its performance, particularly in settings where knowledge of the field tends to be highly judgmental and poorly quantified. We use the term "understanding" here to refer to the degree of judgmental or structural knowledge (as opposed to data) that is contained in the program. Analyses of human clinical decision making [14], [48] suggest that as decisions move from simple to complex, a physician's reasoning style becomes less algorithmic and more heuristic, with qualitative judgmental knowledge and the conditions for invoking it coming increasingly into play. It is likely that medical computing researchers will similarly have to become "knowledge engineers" in the sense that they will look for effective ways to match the knowledge structures that they use to the complexity of the tasks they are undertaking.

Acceptability Issues

A recurring observation as one reviews the literature of computer-based medical decision making is that essentially none of the systems has been effectively utilized outside of a research environment, even when its performance has been shown to be excellent! This suggests that it may be an error to concentrate our research effort primarily on improving the decision making performance of computers when there is evidently much more required before these systems will have clinical impact. It is tempting to conclude that the biases of medical personnel against computers are so strong that systems will inevitably be rejected, regardless of performance, and in fact there are some data to support this view [99]. However, we are beginning to see examples of applications in which initial resistance to automated techniques has gradually been overcome through the incorporation of adequate system benefits [113].

Perhaps one of the most revealing lessons on this subject is an observation regarding the system of Mesel et al. that we described earlier [64]. Despite documented physician resistance to clinical algorithms in other settings [34], the physicians in Mesel's study accepted the guidance of protocols for the management of chemotherapy in their cancer patients. It is likely that the key to acceptance in this instance is the fact that these physicians had previously had no choice but to refer their patients with cancer to the tertiary care center in Birmingham where all complex chemotherapy was administered. The introduction of the protocols permitted these physicians to undertake tasks that they had previously been unable to do, and it simultaneously allowed maintenance of close doctor- patient relationships and helped the patients avoid frequent long trips to the center. The motivation for the physician to use the system is clear in this case. It is reminiscent of Rosati's assertion that physicians will first welcome computer decision aids when they become aware that colleagues who are using the machine have a clear advantage in their practice [81].

A heightened awareness of "human engineering" issues among medical computing researchers is also apt to help improve acceptance of computers by physicians. Fox has recently reviewed this field in detail [18]. The issues range from the mechanics of interaction at a computer terminal to program characteristics designed to make the system appear as a tool for the physician rather than a dogmatic advice-giving machine.

Adequate attention must also be given to the severe time constraints perceived by physicians. Ideally they would like programs to take no more time than they currently spend when accomplishing the same task on their own. Time and schedule pressures are similarly likely to explain the greater resistance to automation among interns and residents than among medical students or practicing physicians in Startsman's study [99].

Finally it must be noted that acceptability issues should generally be considered from the outset in a system's design because they may dictate the choice of methodology as much as the task domain itself does. The role of formal knowledge structures to facilitate explanation capabilities, for example, may argue in favor of using symbolic reasoning techniques even when a somewhat less complex methodology might have been adequate for the decision task.

In summary, the trend towards increased use of knowledge engineering techniques for clinical decision programs has been in response to desires for both improved performance and improved acceptance of such systems. As greater experience is gained with these techniques and they become better known throughout the medical computing community, it is likely that we will see increasingly powerful unions between symbolic reasoning and the alternate methodologies we have discussed. One lesson to be drawn lies in the recognition that there is basic computer science research to be done in medical computing, and that the field is more than the application of established computing techniques in medical domains.

Acknowledgments

We wish to thank R. Blum, L. Fagan, J. King, J. Kunz, H. Sox, and G. Wiederhold. for their thoughtful advice in reviewing earlier drafts of this paper.

References

1. Armitage, P. and Gehan, E.A. "Statistical methods for the identification and use of prognostic factors." Int. J. Cancer, 13, pp. 16-36, (1974).

2. Bleich, H.L. "Computer evaluation of acid-base disorders." J. Clin. Invest. 48, pp. 1689-1696 (1969).

3. Bleich, H.L. "The computer as a consultant." N. Eng. J. Med. 284, pp. 141-147 (1971).

4. Bleich, H.L. "Computer-based consultation: electrolyte and acid-base disorders." Amer. J. Med 53, pp. 285-291 (1972).

5. Blum, R.L. and Wiederhold, G. " Inferring knowledge from clinical data banks: utilizing techniques from artificial intelligence," Proc. 2nd Ann. Symp. on Comp. Applic. in Med. Care, IEEE, Washington D.C., November 1978, pp. 303-307.

6. Buchanan, B.G. and Feigenbaum, E.A. "Dendral and Meta-Dendral: their applications dimension." Artificial Intelligence 11, pp. 5-24 (1978).

7. Croft, D.J. "Is computerized diagnosis possible?" Comp. Biomed. Res. 5, pp. 351-367 (1972).

8. Cumberpatch, J. and Heaps, H.S. "A disease-conscious method for sequential diagnosis by use of disease probabilities without assumption of symptom independence." Int. J. Biomed. Comput. 7, pp. 61-78 (1976).

9. Davis, R. and King, J. "An overview of production systems." In Machine Representation of Knowledge (E.W. Elcock and D. Michie, eds.), New York: Wiley, 1976.

10. deDombal, F.T., Leaper, D.J., Staniland, J.R., et al. "Computer-aided diagnosis of acute abdominal pain." Brit. Med. J. 2, pp.9-13 (1972).

11. deDombal, F.T., Leaper, D.J., Horrocks, J.C., et al. "Human and computer-aided diagnosis of abdominal pain: further report with emphasis on performance of clinicians." Brit. Med. J. 1, pp.376-380 (1974).

12. Duda, R.O. and Hart, P.E. Pattern Classification and Scene Analysis. New York: Wiley, 1973.

13. Edwards, W. "N=1: diagnosis in unique cases." In Computer Diagnosis And Diagnostic Methods, (J.A. Jacquez, ed.), Springfield, Ill.: Charles C. Thomas, 1972, pp. 139-151.

14. Elstein, A.S., Shulman, L.S., and Sprafka, S.A. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, Mass.: Harvard Univ. Press, 1978.

15. Feigenbaum, E.A. "The Art of Artificial Intelligence: Themes and case studies of knowledge engineering." AFIPS Conference Proc., NCC 1978. Vol. 47. Montvale, N.J.: AFIPS Press, 1978, p.227.

16. Feinstein, A.R. "Quality of data in the medical record." Comput. Biomed. Res. 3, pp. 426-435 (1970).

17. Feinstein, A.R., Rubinstein, J.F., and Ramshaw, W.A. "Estimating prognosis with the aid of a conversational mode computer program." Anns. Int. Med. 76, pp. 911-921 (1972).

18. Fox, J. "Medical computing and the user." Int. J. Man-Machine Studies 9, pp. 669-686 (1977).

19. Friedman, R.B. and Gustafson, D.H. "Computers in clinical medicine: a critical review." Comp. Biomed. Res. 8, pp. 199-204 (1977).

20. Fries, J.F. "Time-oriented patient records and a computer databank." J. Amer. Med. Assoc. 222, pp. 1536-1542 (1972).

21. Fries, J.F. "A data bank for the clinician?" (editorial). N. Eng. J. Med. 294, pp. 1400-1402 (1976).

22. Garland, L.H. "Studies on the accuracy of diagnostic procedures." Amer. J. Roentgen. 82, pp. 25-38 (1959).

23. Gill, P.W., Leaper, D.J., Guillou, P.J., et al. "Observer variation in clinical diagnosis - a computer-aided assessment of its magnitude and importance." Meth. Inform. Med.. 12; pp. 108-113 (1973).

24. Ginsberg, A.S. Decision Analysis in Clinical Patient Management With an Application to the Pleural Effusion Syndrome. The Rand Corporation, R-751-RC/NLM, July 1971.

25. Ginsberg, A.S. "The diagnostic process viewed as a decision problem." In Computer Diagnosis and Diagnostic Methods, (J.A. Jacquez, ed.), Springfield, Ill.: Charles C. Thomas, 1972.

26. Gleser, M.A. and Collen, M.F. "Towards automated medical decisions." Comp. Biomed. Res. 5, pp. 180-189 (1972).

27. Goldwyn, R.M., Friedman, H.P., Siegel, J.H. "Iteration and interaction in computer data bank analysis: as case study in the physiologic classification and assessment of the critically ill." Comp. Biomed. Res. 6(1973).

28. Gorry, G.A. and Barnett, G.O. "Experience with a model of sequential diagnosis." Comp. Biomed. Res. 1, pp. 490-507 (1968).

29. Gorry, G.A., Kassirer, J.P., Essig, A., and Schwartz, W.B. "Decision analysis as the basis for computer-aided management of acute renal failure." Amer. J. Med 55, pp. 473-484 (1973).

30. Gorry, G.A. "Computer-assisted clinical decision making." Meth. Inform. Med. 12, pp. 45-51 (1973).

31. Gorry, G.A., Silverman, H., and Pauker, S.G. "Capturing clinical expertise: a computer program that considers clinical responses to digitalis." Amer. J. Med 64, pp. 452-460 (1978).

32. Greenes, R.A., Barnett, G.O., Klein, S.W., et al. "Recording, retrieval,

and review of medical data by physician-computer interaction." N. Eng. J. Med. 282, pp. 307-315 (1970).

33. Greenfield, S., Komaroff, A.L., and Anderson, H. "A headache protocol for nurses: effectiveness and efficiency." Arch. Intern. Med. 136, pp. 1111-1116 (1976).

34. Grimm, R.H., Shimoni, K., Harlan, W.R., and Estes, E.H. "Evaluation of patient-care protocol use by various providers." N. Eng. J. Med. 292, pp. 507-511 (1975).

35. Groner, G.F., Clark, R.L., Berman, R.A., and De Land, E.C. "BIOMOD - an interactive computer graphics system for modeling." Proc. Fall Joint Computer Conference, pp. 369-378, 1971.

36. Hess, E.V. "A uniform database for rheumatic diseases." Arthritis and Rheumatism 19, pp. 645-648 (1976).

37. Hewitt, C. Description and Theoretical Analysis (Using Schemata) of PLANNER: A Language for Proving Theorems and Manipulating Models in a Robot. Ph.D. Dissertation, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Mass., 1972.

38. Horrocks, J.C., McCann, A.P., Staniland, J.R., et al. "Computer-aided diagnosis: description of an adaptable system, and operational experience with 2,034 cases." Brit. Med. J. 2, pp. 5-9 (1972).

39. Horrocks, J.C., and deDombal, F.T. "Computer-aided diagnosis of dyspepsia." Amer. J. Diges. Dis. 20,397-406 (1975).

40. Howard, R. A. (ed.). "Special Issue on Decision Analysis." IEEE Transactions on Systems, Science and Cybernetics, vol SSC-4(3), Sept., 1968.

41. Inglefinger, F.J. "Decision in medicine" (editorial). N. Eng. J. Med. 293, pp. 254-255 (1975).

42. Jacquez, J.A. Computer Diagnosis and Diagnostic Methods, Springfield, Ill.: Charles C. Thomas, 1972.

43. Jelliffe, R.W., Buell, J., Kalaba, R., et al. "A computer program for digitalis dosage regimens." Math. Biosci. 9, pp. 179-193 (1970).

44. Jelliffe, R.W., Buell, J., and Kalaba, R. "Reduction of digitalis toxicity by computer-assisted glycoside dosage regimens." Anns. Int. Med. 77, pp. 891-906 (1972).

45. Johnson, D.C. and Barnett, G.O. "MEDINFO - a medical information system." Comp. Prog. in Biomed. 7, pp. 191-201 (1977).

46. Kanal, L.N. "Patterns in Pattern Recognition: 1968-1974," IEEE Trans. on Information Theory, vol. IT-20, no. 6 (1974).

47. Karpinski, R.H.S. and Bleich, H.L. "MISAR: a miniature information storage and retrieval system." Comp. Biomed. Res. 4, pp. 655-660 (1971).

48. Kassirer, J.P. and Gorry, G.A. "Clinical problem solving: a behavioral analysis." Anns. Int. Med. 89, pp. 245-255 (1978).

49. Kleinmuntz, B. and McLean, R.S. "Diagnostic interviewing by digitalcomputer." Behav. Sci. 13, pp. 75-80 (1968).

50. Knapp, R.G., Levi, S., Lurie, D., and Westphal, M. " A computer-generated diagnostic decision guide: a comparison of statistical diagnosis and clinical diagnosis." Comput. Biol. Med. 7, pp. 223-230 (1977).

51. Komoroff, A.L., Black, W.L., Flatley, M., et al. "Protocols for physician assistants: management of diabetes and hypertension." N. Eng. J. Med. 290,307-312 (1974).

52. Korein, J., Lyman, M., and Tick, J.L. " The computerized medical record," Bulletin New York Academy of Medicine, Vol.47, pp. 824-826 (1971).

53. Koss, N. and Feinstein, A.R. "Computer-aided prognosis: II. development of a prognostic algorithm." Arch. Intern. Med. 127, pp. 448-459 (1971).

54. Leaper, D.J., Horrocks, J.C., Staniland, J.P., and deDombal, F.T. "Computer-assisted diagnosis of abdominal pain using 'estimates' provided by clinicians." Brit. Med. J. 4, pp. 350-354 (1972).

55. Ledley, R.S. and Lusted, L.B. "Reasoning foundations of medical diagnosis." Science 130,9-21 (1959).

56. Levi, S., Frant, J.R., Westphal, M.C., and Lurie, D. "Development of a decision guide - optimal discriminations for meningitis determined by statistical analysis." Meth. Inform. Med. 15 (2), 87-90 (1976).

57. Lipkin, M. and Hardy, J.D. "Mechanical correlation of data in differential diagnosis of hematologic diseases." J. Amer. Med. Assoc. 166, pp. 113-125 (1958).

58. Lusted, L.B. Introduction To Medical Decision Making. Springfield, Ill.: Charles C. Thomas, 1968.

59. Mabry, J.C., Thompson, H.K., Hopwood, M.D., and Baker, W.R. "A prototype data management and analysis system (CLINFO): system description and user experience." In MEDINFO 77, Amsterdam: North-Holland Publishing Co., 1977, pp. 71-75.

60. McDonald, C., Bhargava, B., and Jeris, D. "A clinical information system (CIS) for ambulatory care," Proc. of the 1975 NCC, AFIPS Press, vol. 44 (1975) pp. 749-756

61. McNeil, B.J., Keeler, E., and Adelstein, S.J. "Primer on certain elements of medical decision making." N. Eng. J. Med. 293, pp. 211-215 (1975).

62. McNeil, B.J. and Adelstein, S.J. "Determining the value of diagnostic and screening tests." J. Nucl. Med. 17, pp. 439-448 (1977).

63. Menn, S.J., Barnett, G.O., Schmechel, D., et al. "A computer program to assist in the care of acute respiratory failure." J. Amer. Med. Assoc. 223, pp. 308-312 (1973).

64. Mesel, E., Wirtschafter, D.D., Carpenter, J.T., et al. Clinical algorithms for cancer chemotherapy - systems for community-based consultant-extenders and oncology centers. Meth. Inform. Med. 15, pp. 168-173 (1976).

65. Nordyke, R.A., Kulikowski, C.A., and Kulikowski, C.W. "A comparison of methods for the automated diagnosis of thyroid dysfunction." Comp. Biomed. Res. 4, pp. 374-389 (1971).

66. Norusis, M.J. and Jacquez, J.A. "Diagnosis. I. Symptom nonindependence in mathematical models for diagnosis." Comp. Biomed. Res. 8, pp. 156-172 (1975).

67. Patrick, E.A. "Pattern Recognition in Medicine," Systems, Man and Cybernetics Review, 6, p. 4 (1977).

68. Pauker, S.G. and Kassirer, J.P. "Therapeutic decision making: a cost-benefit analysis." N. Eng. J. Med. 293, pp. 229-234 (1975).

69. Pauker, S.G., Gorry, G.A., Kassirer, J.P., and Schwartz, W.B. "Towards the simulation of clinical cognition: taking a present illness by computer." Amer. J. Med. 60:981-996 (1976).

70. Pauker, S.G. "Coronary artery surgery: the use of decision analysis." Anns. Int. Med. 85, pp. 8-18 (1976).

71. Pauker, S.P. and Pauker, S.G. "Prenatal diagnosis: a directive approach to genetic counseling using decision analysis." Yale J. Biol. Med. 50,275-289 (1977).

72. Peck, C.C., Sheiner, L.B., Martin, C.M., et al. "Computer-assisted digoxin therapy." N. Eng. J. Med. 289, pp. 441-446 (1973).

73. Pipberger, H.V. "Clinical application of a second generation electrocardiography computer program." Amer. J. Electrocardiology 35, pp. 597- 608 (1975).

74. Pliskin, • J.S. and Beck, C.H. "Decision analysis in individual clinical decision making: a real-world application in treatment of renal disease." Meth. Inform. Med. 15, pp. 43-46 (1976).

75. Pople, H.E., Myers, J.D. and Miller, R.A. "DIALOG: A model of diagnostic logic for internal medicine." Proc. 4th Int. Joint. Conf. on Artif. Intell., MIT, Cambridge, Mass., 1975.

76. Pople, H. "The formation of composite hypotheses in diagnostic problem solving: an exercise in synthetic reasoning." Proc. of 5th Intl Joint Conf on Artif. Intelligence, Cambridge, Mass, 1977, pp. 1030-1037.

77. Prutting, J. "Lack of correlation between antemortem and postmortem diagnosis." N.Y. J. Med. 67, pp. 2081-2084 (1967).

78. Raiffa, H. Decision Analysis: Introductory Lectures on Choices Under Uncertainty. Reading, Mass.: Addison Wesley, 1968.

79. Richards, B. and Goh, A.E.S. "Computer assistance in the treatment of patients with acid-base and electrolyte disturbances." MEDINFO 77, Amsterdam: North-Holland Publishing Company, 1977, pp. 407-410.

80. Rodnick, J., and Wiederhold, G. , "Review of automated ambulatory medical record systems: charting services that are of essential benefit to the physician," MEDINFO 77, Amsterdam: North-Holland Publishing Co., 1977, pp. 957-961.

81. Rosati, R.A., Wallace, A.G., and Stead, E.A. "The way of the future." Arch. Intern. Med. 131, pp. 285-287 (1973).

82. Rosati, R.D., McNeer, J.F., Starmer, C.F., et al. "A new information system for medical practice." Arch. Intern. Med. 135, pp. 1017-1024 (1975).

83. Rosenblatt, M.B., Teng, P.K., and Kerpe, S. "Diagnostic accuracy in cancer as determined by post-mortem examination." Prog. Clin. Cancer 5, pp. 71-80 (1973).

84. Rubin, A.D. and Risley, J.F. "The PROPHET system: an experiment in providing a computer resource to scientists." MEDINFO 77, Amsterdam: North-Holland Publishing Co., 1977, pp. 77-81.

85. Safran, C., Tsichlis, P.N., Bluming, A.Z., and Desforges, J.F. "Diagnostic planning using computer-assisted decision making for patients with Hodgkins' disease." Cancer 39, pp. 2426-2434 (1977).

86. Schoolman, H. and Bernstein, L. "Computer use in diagnosis, prognosis, and therapy." Science 200, pp. 926-931 (1978).

87. Schwartz, W.B., Gorry, G.A., Kassirer, J.P., and Essig, A. "Decision analysis and clinical judgment." Amer. J. Med 55, pp. 459-472 (1973).

88. Scott, A.C., Clancey, W., Davis, R., and Shortliffe, E.H. "Explanation capabilities of knowledge-based production systems." Amer. J. Computational Linguistics, Microfiche 62, 1977.

89. Sheiner, L.B., Halkin, H., Peck, C., et al. "Improved computer-assisted digoxin therapy." Anns. Int. Med. 82, pp. 619-627 (1975).

90. Sherman, H., Reiffen, B., and Komoroff, A.L. "Ambulatory care systems." In Problem-Directed and Medical Information Systems (M.F. Driggs, ed.), New York: Intercontinental Medical Book Corporation, 1973, pp. 143-171.

91. Shimura, M. "Learning procedures in pattern classifiers - introduction and survey." Proc. Intl. Joint Conf. on Pattern Recognition, Kyoto, 1978, pp. 125-138.

92. Shortliffe, E.H., Axline, S.G., Buchanan, B.G., and Cohen, S.N. "Design considerations for a program to provide consultations in clinical therapeutics." Proc. 13th San Diego Biomedical Symposium, 311-319, San Diego, Calif., February 1974.

93. Shortliffe, E.H. and Davis, R. "Some considerations for the implementation of knowledge-based expert systems." SIGART Newsletter, No. 55, 9-12, December 1975.

94. Shortliffe, E.H., and Buchanan, B.G. "A model of inexact reasoning in medicine." Math. Biosci. 23, pp. 351-379 (1975).

95. Shortliffe, E.H. Computer-Based Medical Consultations: MYCIN, New York: Elsevier/North Holland, 1976.

96. Slamecka, V., Camp, H.N., Badre, A.N., and Hall, W.D. "MARIS: a knowledge system for internal medicine." Inform. Process & Man. 13, pp. 273-276 (1977).

97. Sox, H.C., Sox, C.H., and Tompkins, R.K. "The training of physicians' assistants: the use of a clinical algorithm system." N. Eng. J. Med. 288, pp. 818-824 (1973).

98. Sridharan, N.S. Guest editorial. Artificial Intelligence 11, pp. 1-4 (1978).

99. Startsman, T.S., and Robinson, R.E. "The attitudes of medical and paramedical personnel towards computers." Comp. Biomed. Res. 5, pp. 218-227 (1972).

100. Stead, W.W., Brame, R.G., Hammond, W.E., et al. "A computerized obstetric medical record." Obstet. & Gyn. 49, pp. 502-509 (1977).

101. Szolovits, P. and Pauker, S.G. "Categorical and probabilistic reasoning in medical diagnosis." Artificial Intelligence 11, pp. 115-144 (1978).

102. Taylor, T.R. "Clinical decision analysis." Meth. Inform. Med.. 15, pp. 216-224 (1976).

103. Vickery, D.M. "Computer support of paramedical personnel: the question of quality control." MEDINFO 74, Amsterdam: North-Holland Publishing Company, 1974, pp. 281-287.

104. Wagner, G., Tautu, P., and Wolber, U. "Problems of medical diagnosis: a bibliography." Meth. Info. Med. 17, pp. 55-74 (1978).

105. Walsh, B.T., Bookhein, W.W., Johnson, R.C., et al. "Recognition of streptococcal pharyngitis in adults." Arch. Int. Med. 135, pp. 1493-1497 (1975).

106. Wardle, A. and Wardle, L. "Computer-aided diagnosis: a review of research." Meth. Info. Med. 17, pp. 15-28 (1978).

107. Warner, H.R., Toronto, A.F., and Veasy, L.G. "Experience with Bayes' Theorem for computer diagnosis of congenital heart disease." Anns. N.Y. Acad. Sci. 115, pp. 558-567 (1964).

108. Warner, H.R. "Experiences with computer-based patient monitoring." Anes. & Analgesia Current Researchers 47, pp. 453-461 (1968).

109. Warner, H.R., Olmsted, C.M., and Rutherford, B.D. "HELP - a program for medical decision-making." Comp. Biomed. Res. 5, pp. 65-74 (1972).

110. Warner, H.R., Rutherford, B.D., and Houtchens, B. "A sequential approach to history taking and diagnosis." Comp. Biomed. Res. 5, pp. 256-262 (1972).

111. Warner, H.R., Morgan, J.D., Pryor, T.A., et al. "HELP - a self-improving system for medical decision making." MEDINFO 74, Amsterdam: North-Holland Publishing Company, 1974.

112. Warner, H.R. Knowledge sectors for logical processing of patient data in the HELP system." Proc. of 2nd. Ann. Symp. on Computer Applications in Medical Care, IEEE, Wasn. D.C.,(1978), pp. 401-404.

113. Watson, R.J. "Medical staff response to a medical information system with

direct physician-computer interface." MEDINFO 74, p. 299-302, Amsterdam: North-Holland Publishing Company, 1974.

114. Wechsler, H. "A fuzzy approach to medical diagnosis." Int. J. Biomed. Comp. 7, pp. 191-203 (1976).

115. Weed, L.L. "Medical records that guide and teach." N. Eng. J. Med. 278, pp. 593-599,652-657 (1968).

116. Weed, L.L. "Problem-oriented medical records." In Problem-Directed and Medical Information Systems (M.F. Driggs, ed.), New York: Intercontinental Medical Book Corporation, 1973.

117. Weiss, S.M., Kulikowski, C.A., Amarel, S. and Safir, A. "A model-based method for computer-aided medical decision-making." Artificial Intelligence 11, pp. 145-172 (1978).

118. Weyl, S., Fries, J., Wiederhold, G., and Germano, F. "A modular self-describing clinical databank system." Comp. Biomed. Res. 8, pp. 279-293 (1975).

119. Wiederhold, G., Fries, J.F., and Weyl, S. "Structured organization of clinical databases," Proc. of the 1975 NCC, AFIPS Press vol. 44 (1975) pp. 479-485

120. Winston, P.H. Artificial Intelligence, Reading, Mass.: Addison-Wesley, 1977.

121. Wortman, P.M. "Medical diagnosis: an information processing approach." Comput. Biomed. Res. 5, pp. 315-328 (1972).

122. Yu, V.L., Fagan, L.M., Wraith, S.M., et al. "Computer-based consultation in antimicrobial selection - a comparative evaluation by experts." Stanford University School of Medicine. Submitted for publication, November 1978.

123. Yu, V.L., Buchanan, B.G., Shortliffe, E.H., et al. "An evaluation of the performance of a computer-based consultant." To appear in Comput. Prog. Biomed., 1979.

124. Zadeh, L.A. "Fuzzy sets." Information and Control 8, pp. 338-353 (1965).

125. Zoltie, N., Horrocks, J.C., and deDombal, F.T. "Computer-assisted diagnosis of dyspepsia - report on transferability of a system, with emphasis on early diagnosis of gastric cancer." Meth. Inform. Med. 16, pp. 89-92 (1977).

Appendix C

THE ART OF ARTIFICIAL INTELLIGENCE:

I. Themes and case studies of knowledge engineering

Edward A. Feigenbaum

Department of Computer Science,
Stanford University,
Stanford, California, 94305.

Abstract

The knowledge engineer practices the art of bringing the principles and tools of AI research to bear on difficult applications problems requiring experts' knowledge for their solution. The technical issues of acquiring this knowledge, representing it, and using it appropriately to construct and explain lines-of-reasoning, are important problems in the design of knowledge-based systems. Various systems that have achieved expert level performance in scientific and medical inference illuminate the art of knowledge engineering and its parent science, Artificial Intelligence.

INTRODUCTION: AN EXAMPLE

This is the first of a pair of papers that will examine emerging themes of knowledge engineering, illustrate them with case studies drawn from the work of the Stanford Heuristic Programming Project, and discuss general issues of knowledge engineering art and practice.

Let me begin with an example new to our workbench: a system called PUFF, the early fruit of a collaboration between our project and a group at the Pacific Medical Center (PMC) in San Francisco.

A physician refers a patient to PMC's pulmonary function testing lab for diagnosis of possible pulmonary function disorder. For one of the tests, the patient inhales and exhales a few times in a tube connected to an instrument/computer combination. The instrument acquires data on flow rates and volumes, the so-called flow-volume loop of the patient's lungs and airways. The computer measures certain parameters of the curve and presents them to the diagnostician (physician or PUFF) for interpretation. The diagnosis is made along these lines: normal or diseased; restricted lung disease or obstructive airways disease or a combination of both; the severity; the likely disease type(s) (e.g. emphysema, bronchitis, etc.); and other factors important for diagnosis.

PUFF is given not only the measured data but also certain items of information from the patient record, e.g. sex, age, number of pack-years of cigarette smoking. The task of the PUFF system is to infer a diagnosis and print it out in English in the normal medical summary form of the interpretation expected by the referring physician.

Everything PUFF knows about pulmonary function diagnosis is contained in (currently) 55 rules of the IF...THEN... form. No textbook of medicine currently records these rules. They constitute the partly-public, partly-private knowledge of an expert pulmonary physiologist at PMC, and were extracted and polished by project engineers working intensively with the expert over a period of time. Here is an example of a PUFF rule (the unexplained acronyms refer to various data measurements):

RULE 31

IF:
1) The severity of obstructive airways disease of the patient is greater than or equal to mild, and
2) The degree of diffusion defect of the patient is greater than or equal to mild, and
3) The tlc(body box)observed/predicted of the patient is greater than or equal to 110 and
4) The observed-predicted difference in rv/tlc of the patient is greater than or equal to 10

THEN:
1) There is strongly suggestive evidence (.9) that the subtype of obstructive airways disease is emphysema, and
2) It is definite (1.0) that "OAD, Diffusion Defect, elevated TLC, and elevated RV together indicate emphysema." is one of the findings.

One hundred cases, carefully chosen to span the variety of disease states with sufficient exemplary information for each, were used to extract the 55 rules. As the knowledge emerged, it was represented in rule form, added to the system and tested by running additional cases. The expert was sometimes surprised, sometimes frustrated, by the occasional gaps and inconsistencies in the knowledge, and the incorrect diagnoses that were logical consequences of the existing rule set. The interplay between knowledge engineer and expert gradually expanded the set of rules to remove most of these problems.

As cumulation of techniques in the art demands and allows, a new tool was not invented when an old one would do. The knowledge engineers pulled out of their toolkit a version of the MYCIN system (to be discussed later), with the rules about infectious diseases removed, and used it as the inference engine for the PUFF diagnoses. Thus PUFF, like MYCIN, is a relatively simple backward-chaining inference system. It seeks a valid line-of-reasoning based on its rules and rooted in the instrument and patient data. With a little more work at fitting some existing tools together, PUFF will be able to explain this line-of-reasoning, just as MYCIN does.

As it is, PUFF only prints out the final interpretation, of which the following is an example:

```
PATIENT DATA:
The degree of dyspnea:  MODERATELY-SEVERE
The severity of coughing:  MILD
Sputum production  MODERATELY-SEVERE
The number of pack-years of smoking:  48
referral diagnosis:  BRONCHITIS
IVC/IVC-predicted:  80
RV/RV-predicted:  191
FVC/FVC-predicted:  87
TLC(body box)observed/predicted:  127
Predicted FEV1/FVC:  83
TLC(DLCO)observed/predicted:   83
FEV1/FVC ratio:  50
RV/TLC Observed-Predicted:  21
MMF/MMF-predicted:  19
the DLCO/DLCO-predicted:  48
The slope (F50_obs-F25_obs)/FVC_obs:  19

DEGREE OF OBSTRUCTIVE AIRWAYS DISEASE:

OAD degree by SLOPE: (MODERATELY-SEVERE
700)
OAD degree by MMF:   (SEVERE 900)
OAD degree by FEV1:  (MODERATELY-SEVERE
700)

FINAL OAD DEGREE:    (MODERATELY-SEVERE
910) (SEVERE 900)
No conflict.  Final degree:
(MODERATELY-SEVERE 910)

INTERPRETATION:


Obstruction is indicated by curvature of
the flow-volume loop.
Forced Vital Capacity is normal and peak
flow rates are reduced, suggesting
airway obstruction.
Flow rate from 25-75 of expired volume is
reduced, indicating severe airway
obstruction.
OAD, Diffusion Defect, elevated TLC, and
elevated RV together indicate emphysema.
OAD, Diffusion Defect, and elevated RV
indicate emphysema.
Change in expired flow rates following
bronchodilation shows that there is
reversibility of airway obstruction.
The presence of a productive cough is an
indication that the OAD is of the
bronchitic type.
Elevated lung volumes indicate
overinflation.
Air trapping is indicated by the elevated
difference between observed and predicted
RV/TLC ratios.
Improvement in airway resistance indicates
some reversibility of airway
Airway obstruction is consistent with the
patient's smoking history.
The airway obstruction accounts for the
patient's dyspnea.
Although bronchodilators were not
useful in this one case, prolonged use may
prove to be beneficial to the patient.
The reduced diffusion capacity indicates
airway obstruction of the mixed
bronchitic and emphysematous types.
Low diffusing capacity indicates loss of
alveolar capillary surface.
Obstructive Airways Disease of mixed types
```

150 cases not studied during the knowledge acquisition process were used for a test and validation of the rule set. PUFF inferred a diagnosis for each. PUFF-produced and expert-produced interpretations were coded for statistical analysis to discover the degree of agreement. Over various types of disease states, and for two conditions of match between human and computer diagnoses ("same degree of severity" and "within one degree of severity"), agreement ranged between approximately 90% and 100%.

The PUFF story is just beginning and will be told perhaps at the next IJCAI. The surprising punchline to my synopsis is that the current state of the PUFF system as described above was achieved in less than 50 hours of interaction with the expert and less than 10 man-weeks of effort by the knowledge engineers. We have learned much in the p.ıst decade of the art of engineering knowledge-based intelligent agents!

In the remainder of this essay, I would like to discuss the route that one research group, the Stanford Heuristic Programming Project, has taken, illustrating progress with case studies, and discussing themes of the work.

2 ARTIFICIAL INTELLIGENCE & KNOWLEDGE ENGINEERING

The dichotomy that was used to classify the collected papers in the volume Computers and Thought still characterizes well the motivations and research efforts of the AI community. First, there are some who work toward the construction of intelligent artifacts, or seek to uncover principles, methods, and techniques useful in such construction. Second, there are those who view artificial intelligence as (to use Newell's phrase) "theoretical psychology," seeking explicit and valid information processing models of human thought.

For purposes of this essay, I wish to focus on the motivations of the first group, these days by far the larger of the two. I label these motivations "the intelligent agent viewpoint" and here is my understanding of that viewpoint:

> "The potential uses of computers by people to accomplish tasks can be 'one-dimensionalized' into a spectrum representing the nature of instruction that must be given the computer to do its job. Call it the WHAT-TO-HOW spectrum. At one extreme of the spectrum, the user supplies his intelligence to instruct the machine with precision exactly HOW to do his job, step-by-step. Progress in Computer Science can be seen as steps away from the extreme 'HOW' point on the spectrum: the familiar panoply of assembly languages, subroutine libraries, compilers, extensible languages, etc. At the other extreme of the spectrum is the user with his real problem (WHAT he wishes the computer, as his instrument, to do for him). He aspires to communicate WHAT he wants done in a language that is comfortable to him (perhaps English); via communication modes that are convenient for him (including perhaps, speech or pictures); with some generality, some vagueness, imprecision, even error; without having to lay out in detail all necessary subgoals for adequate performance - with reasonable assurance that he is addressing an intelligent agent that is using knowledge of his world to understand his intent, to fill in his vagueness, to make specific his abstractions, to correct his errors, to discover appropriate subgoals, and ultimately to translate WHAT he really wants done into processing steps that define HOW it shall be done by a real computer. The research activity aimed at creating computer programs that act as "intelligent agents" near the WHAT end of the WHAT-To-HOW spectrum can be viewed as the long-range goal of AI research." (Feigenbaum, 1974)

Our young science is still more art than science. Art: "the principles or methods governing any craft or branch of learning." Art: "skilled workmanship, execution, or agency." These the dictionary teaches us. Knuth tells us that the endeavor of computer programming is an art, in just these ways. The art of constructing intelligent agents is both part of and an extension of the programming art. It is the art of building complex computer programs that represent and reason with knowledge of the world. Our art therefore lives in symbiosis with the other worldly arts, whose practitioners -- experts of their art — hold the knowledge we need to construct intelligent agents. In most "crafts or branches of learning" what we call "expertise" is the essence of the art. And for the domains of knowledge that we touch with our art, it is the "rules of expertise" or the rules of "good judgment" of the expert practitioners of that domain that we seek to transfer to our programs.

2.1 Lessons of the Past

Two insights from previous work are pertinent to this essay.

The first concerns the quest for generality and power of the inference engine used in the performance of intelligent acts (what Minsky and Papert [see Goldstein and Papert, 1977] have labeled "the power strategy"). We must hypothesize from our experience to date that the problem solving power exhibited in an intelligent agent's performance is primarily a consequence of the specialist's knowledge employed by the agent, and only very secondarily related to the generality and power of the inference method employed. Our agents must be knowledge-rich, even if they are methods-poor. In 1970, reporting the first major summary-of-results of the DENDRAL program (to be discussed later), we addressed this issue as follows:

> "...general problem-solvers are too weak to be used as the basis for building high-performance systems. The behavior of the best general problem-solvers we know, human problem-solvers, is observed to be weak and shallow, except in the areas in which the human problem-solver is a specialist. And it is observed that the transfer of expertise between specialty

> areas is slight. A chess master is unlikely to be an expert algebraist or an expert mass spectrum analyst, etc. In this view, the expert is the specialist, with a specialist's knowledge of his area and a specialist's methods and heuristics." (Feigenbaum, Buchanan and Lederberg, 1971, p. 187)

Subsequent evidence from our laboratory and all others has only confirmed this belief.

AI researchers have dramatically shifted their view on generality and power in the past decade. In 1967, the canonical question about the DENDRAL program was: "It sounds like good chemistry, but what does it have to do with AI?" In 1977, Goldstein and Papert write of a paradigm shift in AI:

> "Today there has been a shift in paradigm. The fundamental problem of understanding intelligence is not the identification of a few powerful techniques, but rather the question of how to represent large amounts of knowledge in a fashion that permits their effective use and interaction." (Goldstein and Papert, 1977)

The second insight from past work concerns the nature of the knowledge that an expert brings to the performance of a task. Experience has shown us that this knowledge is largely heuristic knowledge, experiential, uncertain -- mostly "good guesses" and "good practice," in lieu of facts and rigor. Experience has also taught us that much of this knowledge is private to the expert, not because he is unwilling to share publicly how he performs, but because he is unable. He knows more than he is aware of knowing. [Why else is the Ph.D. or the Internship a guild-like apprenticeship to a presumed "master of the craft?" What the masters really know is not written in the textbooks of the masters.] But we have learned also that this private knowledge can be uncovered by the careful, painstaking analysis of a second party, or sometimes by the expert himself, operating in the context of a large number of highly specific performance problems. Finally, we have learned that expertise is multi-faceted, that the expert brings to bear many and varied sources of knowledge in performance. The approach to capturing his expertise must proceed on many fronts simultaneously.

2.2 The Knowledge Engineer

The knowledge engineer is that second party just discussed. [An historical note about the term. In the mid-60s, John McCarthy, for reasons obvious from his work, had been describing Artificial Intelligence as "Applied Epistemology." When I first described the DENDRAL program to Donald Michie in 1968, he remarked that it was "epistemological engineering," a clever but ponderous and unpronounceable turn-of-phrase that I simplified into "knowledge engineering."] She (in deference to my favorite knowledge engineer) works intensively with an expert to acquire domain-specific knowledge and organize it for use by a program. Simultaneously she is matching the tools of the AI workbench to the task at hand -- program organizations, methods of symbolic inference, techniques for the structuring of symbolic information, and the like. If the tool fits, or nearly fits, she uses it. If not, necessity mothers AI invention, and a new tool gets created. She builds the early versions of the intelligent agent, guided always by her intent that the program eventually achieve expert levels of performance in the task. She refines or reconceptualizes the system as the increasing amount of acquired knowledge causes the AI tool to "break" or slow down intolerably. She also refines the human interface to the intelligent agent with several aims: to make the system appear "comfortable" to the human user in his linguistic transactions with it; to make the system's inference processes understandable to the user; and to make the assistance controllable by the user when, in the context of a real problem, he has an insight that previously was not elicited and therefore not incorporated.

In the next section, I wish to explore (in summary form) some case studies of the knowledge engineer's art.

3 CASES FROM THE KNOWLEDGE ENGINEER'S WORKSHOP

I will draw material for this section from the work of my group at Stanford. Much exciting work in knowledge engineering is going on elsewhere. Since my intent is not to survey literature but to illustrate themes, at the risk of appearing parochial I have used as case studies the work I know best.

My collaborators (Professors Lederberg and Buchanan) and I began a series of projects, initially the development of the DENDRAL program, in 1965. We had dual motives: first, to study scientific problem solving and discovery, particularly the processes scientists do use or should use in inferring hypotheses and theories from empirical evidence; and second, to conduct this study in such a way that our experimental programs would one day be of use to working scientists, providing intelligent assistance on important and difficult problems. By 1970, we and our co-workers had gained enough experience that we felt comfortable in laying out a program of research encompassing work on theory formation, knowledge utilization, knowledge acquisition, explanation, and knowledge engineering techniques. Although there were some surprises along the way (like the AM program), the general lines of the research are proceeding as envisioned.

THEMES

As a road map to these case studies, it is useful to keep in mind certain major themes:

Generation-and-test: Omnipresent in our experiments is the "classical" generation-and-test framework that has been the hallmark of AI programs for two decades. This is not a consequence of a doctrinaire attitude on our part about heuristic search, but rather of the usefulness and sufficiency of the concept.

Situation => Action Rules: We have chosen to represent the knowledge of experts in this form. Making no doctrinaire claims for the universal applicability of this representation, we nonetheless point to the demonstrated utility of the rule-based representation. From this representation flow rather directly many of the characteristics of our programs: for example, ease of modification of the knowledge, ease of explanation. The essence of our approach is that a rule must capture a "chunk" of domain knowledge that is meaningful, in and of itself, to the domain specialist. Thus our rules bear only a historical relationship to the production rules used by Newell and Simon (1972) which we view as "machine-language programming" of a recognize => act machine.

The Domain-Specific Knowledge: It plays a critical role in organizing and constraining search. The theme is that in the knowledge is the power. The interesting action arises from the knowledge base, not the inference engine. We use knowledge in rule form (discussed above), in the form of inferentially-rich models based on theory, and in the form of tableaus of symbolic data and relationships (i.e. frame-like structures). System processes are made to conform to natural and convenient representations of the domain-specific knowledge.

Flexibility to modify the knowledge base: If the so-called "grain size" of the knowledge representation is chosen properly (i.e. small enough to be comprehensible but large enough to be meaningful to the domain specialist), then the rule-based approach allows great flexibility for adding, removing, or changing knowledge in the system.

Line-of-reasoning: A central organizing principle in the design of knowledge-based intelligent agents is the maintenance of a line-of-reasoning that is comprehensible to the domain specialist. This principle is, of course, not a logical necessity, but seems to us to be an engineering principle of major importance.

Multiple Sources of Knowledge: The formation and maintenance (support) of the line-of-reasoning usually require the integration of many disparate sources of knowledge. The representational and inferential problems in achieving a smooth and effective integration are formidable engineering problems.

Explanation: The ability to explain the line-of-reasoning in a language convenient to the user is necessary for application and for system development (e.g. for debugging and for extending the knowledge base). Once again, this is an engineering principle, but very important. What constitutes "an explanation" is not a simple concept, and considerable thought needs to be given, in each case, to the structuring of explanations.

CASE STUDIES

In this section I will try to illustrate these themes with various case studies.

3.1 DENDRAL: Inferring Chemical Structures

3.1.1 Historical Note

Begun in 1965, this collaborative project with the Stanford Mass Spectrometry Laboratory has become one of the longest-lived continuous efforts in the history of AI (a fact that in no small way has contributed to its success). The basic framework of generation-and-test and rule-based representation has proved rugged and extendable. For us the DENDRAL system has been a fountain of ideas, many of which have found their way, highly metamorphosed, into our other projects. For example, our long-standing commitment to rule-based representations arose out of our (successful) attempt to head off the imminent ossification of DENDRAL caused by the rapid accumulation of new knowledge in the system around 1967.

3.1.2 Task

To enumerate plausible structures (atom-bond graphs) for organic molecules, given two kinds of information: analytic instrument data from a mass spectrometer and a nuclear magnetic resonance spectrometer; and user-supplied constraints on the answers, derived from any other source of knowledge (instrumental or contextual) available to the user.

3.1.3 Representations

Chemical structures are represented as node-link graphs of atoms (nodes) and bonds (links). Constraints on search are represented as subgraphs (atomic configurations) to be denied or preferred. The empirical theory of mass spectrometry is represented by a set of rules of the general form:

```
Situation: Particular atomic
           configuration
           (subgraph)

                |
                | Probability, P,
                |  of occurring
                |
                V

Action:  Fragmentation of the
         particular configuration
         (breaking links)
```

Rules of this form are natural and expressive to mass spectrometrists.

3.1.4 Sketch of Method

DENDRAL's inference procedure is a heuristic search that takes place in three stages, without feedback: plan-generate-test.

"Generate" (a program called CONGEN) is a generation process for plausible structures. Its foundation is a combinatorial algorithm (with mathematically proven properties of completeness and non-redundant generation) that can produce all the topologically legal candidate structures. Constraints supplied by the user or by the "Plan" process prune and steer the generation to produce the plausible set (i.e. those satisfying the constraints) and not the enormous legal set.

"Test" refines the evaluation of plausibility, discarding less worthy candidates and rank-ordering the remainder for examination by the user. "Test" first produces a "predicted" set of instrument data for each plausible candidate, using the rules described. It then evaluates the worth of each candidate by comparing its predicted data with the actual input data. The evaluation is based on heuristic criteria of goodness-of-fit. Thus, "test" selects the "best" explanations of the data.

"Plan" produces direct (i.e. not chained) inference about likely substructure in the molecule from patterns in the data that are indicative of the presence of the substructure. (Patterns in the data trigger the left-hand-sides of substructure rules). Though composed of many atoms whose interconnections are given, the substructure can be manipulated as atom-like by "generate." Aggregating many units entering into a combinatorial process into fewer higher-level units reduces the size of the combinatorial search space. "Plan" sets up the search space so as to be relevent to the input data. "Generate is the inference tactician; "Plan" is the inference strategist. There is a separate "Plan" package for each type of instrument data, but each package passes substructures (subgraphs) to "Generate." Thus, there is a uniform interface between "Plan" and "Generate." User-supplied constraints enter this interface, directly or from user-assist packages, in the form of substructures.

3.1.5 Sources of Knowledge

The various sources of knowledge used by the DENDRAL system are:

Valences (legal connections of atoms); stable and unstable configurations of atoms; rules for mass spectrometry fragmentations; rules for NMR shifts; expert's rules for planning and evaluation; user-supplied constraints (contextual).

3.1.6 Results

DENDRAL's structure elucidation abilities are, paradoxically, both very general and very narrow. In general, DENDRAL handles all molecules, cyclic and tree-like. In pure structure elucidation under constraints (without instrument data),CONGEN is unrivaled by human performance. In structure elucidation with instrument data, DENDRAL's performance rivals expert human performance only for a small number of molecular families for which the program has been given specialist's knowledge, namely the families of interest to our chemist collaborators. I will spare this computer science audience the list of names of these families. Within these areas of knowledge-intensive specialization, DENDRAL's performance is usually not only much faster but also more accurate than expert human performance.

The statement just made summarizes thousands of runs of DENDRAL on problems of interest to our experts, their colleagues, and their students. The results obtained, along with the knowledge that had to be given to DENDRAL to obtain them, are published in major journals of chemistry. To date, 25 papers have been published there, under a series title "Applications of Artificial Intelligence for Chemical Inference: <specific subject>" (see references).

The DENDRAL system is in everyday use by Stanford chemists, their collaborators at other universities and collaborating or otherwise interested chemists in industry. Users outside

Stanford access the system over commercial computer/communications network. The problems they are solving are often difficult and novel. The British government is currently supporting work at Edinburgh aimed at transferring DENDRAL to industrial user communities in the UK.

3.1.7 Discussion

Representation and extensibility. The representation chosen for the molecules, constraints, and rules of instrument data interpretation is sufficiently close to that used by chemists in thinking about structure elucidation that the knowledge base has been extended smoothly and easily, mostly by chemists themselves in recent years. Only one major reprogramming effort took place in the last 9 years -- when a new generator was created to deal with cyclic structures.

Representation and the Integration of multiple sources of knowledge. The generally difficult problem of integrating various sources of knowledge has been made easy in DENDRAL by careful engineering of the representations of objects, constraints, and rules. We insisted on a common language of compatibility of the representations with each other and with the inference processes: the language of molecular structure expressed as graphs. This leads to a straightforward procedure for adding a new source of knowledge, say, for example, the knowledge associated with a new type of instrument data. The procedure is this: write rules that describe the effect of the physical processes of the instrument on molecules using the situation => action form with molecular graphs on both sides; any special inference process using these rules must pass its results to the generator only(!) in the common graph language.

It is today widely believed in AI that the use of many diverse sources of knowledge in problem solving and data interpretation has a strong effect on quality of performance. How strong is, of course, domain-dependent, but the impact of bringing just one additional source of knowledge to bear on a problem can be startling. In one difficult (but not unusually difficult) mass spectrum analysis problem*, the program using its mass spectrometry knowledge alone would have generated an impossibly large set of plausible candidates (over 1.25 million!). Our engineering response to this was to add another source of data and knowledge, proton NMR. The addition on a simple interpretive theory of this NMR data, from which the program could infer a few additional constraints, reduced the set of plausible candidates to one, the right structure! This was not an isolated result but showed up dozens of times in subsequent analyses.

DENDRAL and data. DENDRAL's robust models (topological, chemical, instrumental) permit a strategy of finding solutions by generating hypothetical "correct answers" and choosing among these with critical tests. This strategy is opposite to that of piecing together the implications of each data point to form a hypothesis. We call DENDRAL's strategy largely model-driven, and the other data-driven. The consequence of having enough knowledge to do model-driven analysis is a large reduction in the amount of data that must be examined since data is being used mostly for verification of possible answers. In a typical DENDRAL mass spectrum analysis, usually no more than about 25 data points out of a typical total of 250 points are processed. This important point about data reduction and focus-of-attention has been discussed before by Gregory (1968) and by the vision and speech research groups, but is not widely understood.

Conclusion. DENDRAL was an early herald of AI's shift to the knowledge-based paradigm. It demonstrated the point of the primacy of domain-specific knowledge in achieving expert levels of performance. Its development brought to the surface important problems of knowledge representation, acquisition, and use. It showed that, by and large, the AI tools of the first decade were sufficient to cope with the demands of a complex scientific problem-solving task,or were readily extended to handle unforseen difficulties. It demonstrated that AI's conceptual and programming tools were capable of producing programs of applications interest, albeit in narrow specialties. Such a demonstration of competence and sufficiency was important for the credibility of the AI field at a critical juncture in its history.

3.2 META-DENDRAL: inferring rules of mass spectrometry

3.2.1 Historical note

The META-DENDRAL program is a case study in automatic acquisition of domain knowledge. It arose out of our DENDRAL work for two reasons: first, a decision that with DENDRAL we had a sufficiently firm foundation on which to pursue our long-standing interest in processes of scientific theory formation; second, by a recognition that the acquisition of domain knowledge was the bottleneck problem in the building of applications-oriented intelligent agents.

3.2.2 Task

META-DENDRAL's job is to infer rules of fragmentation of molecules in a mass spectrometer for possible later use by the DENDRAL performance

----

* the analysis of an acyclic amine with formula $C_{20}H_{45}N$.

program. The inference is to be made from actual spectra recorded from known molecular structures. The output of the system is the set of fragmentation rules discovered, summary of the evidence supporting each rule, and a summary of contra-indicating evidence. User-supplied constraints can also be input to force the form of rules along desired lines.

3.2.3 Representations

The rules are, of course, of the same form as used by DENDRAL that was described earlier.

3.2.4 Sketch of Method

META-DENDRAL, like DENDRAL, uses the generation-and-test framework. The process is organized in three stages: Reinterpret the data and summarize evidence (INTSUM); generate plausible candidates for rules (RULEGEN); test and refine the set of plausible rules (RULEMOD).

INTSUM: gives every data point in every spectrum an interpretation as a possible (highly specific) fragmentation. It then summarizes statistically the "weight of evidence" for fragmentations and for atomic configurations that cause these fragmentations. Thus, the job of INTSUM is to translate data to DENDRAL subgraphs and bond-breaks, and to summarize the evidence accordingly.

RULEGEN: conducts a heuristic search of the space of all rules that are legal under the DENDRAL rule syntax and the user-supplied constraints. It searches for plausible rules, i.e. those for which positive evidence exists. A search path is pruned when there is no evidence for rules of the class just generated. The search tree begins with the (single) most general rule (loosely put, "anything" fragments from "anything") and proceeds level-by-level toward more detailed specifications of the "anything." The heuristic stopping criterion measures whether a rule being generated has become too specific, in particular whether it is applicable to too few molecules of the input set. Similarly there is a criterion for deciding whether an emerging rule is too general. Thus, the output of RULEGEN is a set of candidate rules for which there is positive evidence.

RULEMOD: tests the candidate rule set using more complex criteria, including the presence of negative evidence. It removes redundancies in the candidate rule set; merges rules that are supported by the same evidence; tries further specialization of candidates to remove negative evidence; and tries further generalization that preserves positive evidence.

3.2.5 Results

META-DENDRAL produces rule sets that rival in quality those produced by our collaborating experts. In some tests, META-DENDRAL recreated rule sets that we had previously acquired from our experts during the DENDRAL project. In a more stringent test involving members of a family of complex ringed molecules for which the mass spectral theory had not been completely worked out by chemists, META-DENDRAL discovered rule sets for each subfamily. The rules were judged by experts to be excellent and a paper describing them was recently published in a major chemical journal (Buchanan, Smith, et al, 1976).

In a test of the generality of the approach, a version of the META-DENDRAL program is currently being applied to the discovery of rules for the analysis of nuclear magnetic resonance data.

3.3 MYCIN and TEIRESIAS: Medical Diagnosis

3.3.1 Historical note

MYCIN originated in the Ph.D. thesis of E. Shortliffe (now Shortliffe, M.D. as well), in collaboration with the Infectious Disease group at the Stanford Medical School (Shortliffe, 1976). TEIRESIAS, the Ph.D. thesis work of R. Davis, arose from issues and problems indicated by the MYCIN project but generalized by Davis beyond the bounds of medical diagnosis applications (Davis, 1976). Other MYCIN-related theses are in progress.

3.3.2 Tasks

The MYCIN performance task is diagnosis of blood infections and meningitis infections and the recommendation of drug treatment. MYCIN conducts a consultation (in English) with a physician-user about a patient case, constructing lines-of-reasoning leading to the diagnosis and treatment plan.

The TEIRESIAS knowledge acquisition task can be described as follows:

In the context of a particular consultation, confront the expert with a diagnosis with which he does not agree. Lead him systematically back through the line-of-reasoning that produced the diagnosis to the point at which he indicates the analysis went awry. Interact with the expert to modify offending rules or to acquire new rules. Rerun the consultation to test the solution and gain the expert's concurrence.

3.3.3 Representations:

MYCIN's rules are of the form:

IF <conjunctive clauses> THEN <implication>

Here is an example of a MYCIN rule for blood infections.

RULE 85

IF:
1) The site of the culture is blood, and
2) The gram stain of the organism is gramneg, and
3) The morphology of the organism is rod, and
4) The patient is a compromised host

THEN:
There is suggestive evidence (.6) that the identity of the organism is pseudomonas-aeruginosa

TEIRESIAS allows the representation of MYCIN-like rules governing the use of other rules,i.e. rule-based strategies. An example follows.

METARULE 2

IF:
1) the patient is a compromised host, and
2) there are rules which mention in their premise pseudomonas
3) there are rules which mention in their premise klebsiellas

THEN:
There is suggestive evidence (.4) that the former should be done before the latter.

3.3.4 Sketch of method

MYCIN employs a generation-and-test procedure of a familiar sort. The generation of steps in the line-of-reasoning is accomplished by backward chaining of the rules. An IF-side clause is either immediately true or false (as determined by patient or test data entered by the physician in the consultation); or is to be decided by subgoaling. Thus, "test" is interleaved with "generation" and serves to prune out incorrect lines-of-reasoning.

Each rule supplied by an expert has associated with it a "degree of certainty" representing the expert's confidence in the validity of the rule (a number from 1 to 10). MYCIN uses a particular ad-hoc but simple model of inexact reasoning to cumulate the degrees of certainty of the rules used in an inference chain (Shortliffe and Buchanan, 1975).

It follows that there may be a number of "somewhat true" lines-of-reasoning -- some indicating one diagnosis, some indicating another. All (above a threshold) are used by the system as sources of knowledge indicating plausible lines-of-reasoning.

TEIRESIAS' rule acquisition process is based on a record of MYCIN's search. Rule acquisition is guided by a set of rule models that dictate the form and indicate the likely content of new rules. Rule models are not given in advance, but are inferred from the knowledge base of existing rules.

3.3.5 Language of Interaction

The language used looks like it might be English but is actually the dialect "Doctor-ese" used by members of the tribe of healing arts practitioners. Reasonably simple language processing methods suffice. When ambiguities of interpretation are encountered, they are fed back to the user for decisions about meaning.

3.3.6 Example diagnosis and treatment plan

The following is an example of a diagnosis reported by MYCIN to one of its users following a consultation. An example consultation will not be shown, but one can be found in Shortliffe's book (Shortliffe, 1976).

My therapy recommendations will be designed to treat for organisms that are either very likely or, although less likely, would have a significant effect on therapy selection if they were present. It is important to

```
cover for the following probable
infection(s) and associated organism(s):


INFECTION-1 is CYSTITIS
<Item 1> PSEUDOMONAS-CEPACIA [ORGANISM-5]
<Item 2> CITROBACTER-DIVERSUS [ORGANISM-4]
<Item 3> E.COLI [ORGANISM-3]

INFECTION-2 is UPPER-RESPIRATORY-INFECTION
<Item 4> LISTERIA  [ORGANISM-2]

INFECTION-3 is BACTEREMIA
<Item 5> ENTEROCOCCUS [ORGANISM-1]

[REC-1] My preferred therapy recommendation
is as follows:

In order to cover for Items <1 2 3 4 5>:
Give the following in combination:
1) KANAMYCIN
Dose: 750 mg (7.5 mg/kg) q12h IM (or IV)
      for 28 days
Comments: Modify dose in renal failure
2) PENICILLIN
Dose: 2,500,000 units (25000 units/kg)
      q4h IV for 28 days

--------------------------------------------------
```

3.3.7 Results

MYCIN is currently being prepared for clinical evaluation. In a preliminary evaluation, done outside of the clinical setting, a panel of expert judges was asked to rate MYCIN's performance along a number of dimensions (e.g. infecting organism identification, organism significance). In 90% of the cases submitted to the judges, a majority of the judges said that the program's decisions were the-same-as or as-good-as the decisions they would have made.

3.3.8 Explanation

MYCIN can expose its line of reasoning by offering explanations of various kinds. "Why" and "How" questions can be asked by the user during the consultation. "Why" is the user's way of inquiring, "Why do you want to know that?" "How" is his way of asking, "How did you conclude that...?" In a post-consultation, he can similarly inquire about how the diagnostic conclusions were reached, and inquire about the reasons why certain other conclusions were ruled out. Examples of the MYCIN-TEIRESIAS explanation capabilities follow:

```
--------------------------------------------------


MYCIN:       Is the patient's illness with
ORGANISM-1 a hospital-acquired infection?
USER: WHY
TEIRESIAS: (i.e. WHY is it important to
determine whether or not the infection with
ORGANISM-1 was acquired while the patient
was hospitalized?)

[1.0] This will aid in determining the
category of ORGANISM-1.



It has already been established that

[1.1] the gram stain of ORGANISM-1 is
gramneg, and
[1.2] the morphology of ORGANISM-1 is rod,
and
[1.3] the aerobicity of ORGANISM-1 is
facultative

Therefore, if

[1.4] the infection with ORGANISM-1 was not
acquired while the patient was hospitalized

then

there is strongly suggestive evidence (.8)
that the category of ORGANISM-1 is
enterobacteriaceae

Also: there is weakly suggestive evidence
(.1) that the identity of
ORGANISM-1 is pseudomonas-aeruginosa
[RULE 37]



USER:  HOW 1.3

TEIRESIAS:  [i.e. HOW was it established
that the aerobicity of ORGANISM-1 is
facultative?]

The following were used:

[3.1] RULE 27 indicated there is strongly
suggestive evidence (.8) that the
aerobicity of ORGANISM-1 is facultative

[3.2] RULE 27 indicated there is weakly
suggestive evidence (.2) that the
aerobicity of ORGANISM-1 is anaerobic

Since this gave a cumulative CF of (.8) for
facultative, and (.2) for anaerobic, it has
been established that the aerobicity of
ORGANISM-1 is facultative.
```

USER: WHY DIDN'T YOU CONSIDER STREPTOCOCCUS AS A POSSIBILITY FOR ORGANISM-1?

TEIRESIAS: The following rule could have been used to determine that the identify of ORGANISM-1 was streptococcus: RULE 33

But clause 2 ("the morphology of the organism is coccus") was already known to be false for ORGANISM-1, so the rule was never tried.

3.3.9 Discussion

A few short comments aimed at knowledge engineering are in order.

English is the language most natural and comfortable for our experts and users, and was therefore the language chosen for interactive consultation, explanation, and external representation of the rules (the internal format is INTERLISP). This situation is not peculiar to doctors; in most areas of application of intelligent agents I believe that English (i.e. natural language) will be the language of choice. Programming an English language processor and front-end to such systems is not a scary enterprise because:

a) the domain is specialized, so that possible interpretations are constrained.

b) specialist-talk is replete with standard jargon and stereotyped ways of expressing knowledge and queries — just right for text templates, simple grammars and other simple processing schemes.

c) the ambiguity of interpretation resulting from simple schemes can be dealt with easily by feeding back interpretations for confirmation. If this is done with a pleasant "I didn't quite understand you..." tone, it is not irritating to the user.

English may be exactly the wrong language for representation and interaction in some domains. It would be awkward, to say the least, to represent DENDRAL's chemical structures and knowledge of mass spectrometry in English, or to interact about these with a user.

Simple explanation schemes have been a part of the AI scene for a number of years and are not hard to implement. Really good models of what explanation is as a transaction between user and agent, with programs to implement these models, will be the subject (I predict) of much future research in AI.

Without the explanation capability, I assert, user acceptance of MYCIN would have been nil, and there would have been a greatly diminished effectiveness and contribution of our experts.

MYCIN was the first of our programs that forced us to deal with what we had always understood: that experts' knowledge is uncertain and that our inference engines had to be made to reason with this uncertainty. It is less important that the inexact reasoning scheme be formal, rigorous, and uniform than it is for the scheme to be natural to and easily understandable by the experts and users.

All of these points can be summarized by saying that MYCIN and its TEIRESIAS adjunct are experiments in the design of a see-through system, whose representations and processes are almost transparently clear to the domain specialist. "Almost" here is equivalent to "with a few minutes of introductory description." The various pieces of MYCIN — the backward chaining, the English transactions, the explanations, etc. — are each simple in concept and realization. But there are great virtues to simplicity in system design; and viewed as a total intelligent agent system, MYCIN/TEIRESIAS' is one of the best engineered.

3.4 SU/X: signal understanding

3.4.1 Historical note

SU/X is a system design that was tested in an application whose details are classified. Because of this, the ensuing discussion will appear considerably less concrete and tangible than the preceding case studies. This system design was done by H.P. Nii and me, and was strongly influenced by the CMU Hearsay II system design.

3.4.2 Task

SU/X's task is the formation and continual updating, over long periods of time, of hypotheses about the identity, location, and velocity of objects in a physical space. The output desired is a display of the "current best hypotheses" with full explanation of the support for each. There are two types of input data: the primary signal (to be understood); and auxiliary symbolic data (to supply context for the understanding). The primary signals are spectra, represented as descriptions of the spectral lines. The various spectra cover the physical space with some spatial overlap.

3.4.3 Representations

The rules given by the expert about objects, their behavior, and the interpretation of signal data from them are all represented in the situation => action form. The "situations" constitute invoking conditions and the "actions" are processes that modify the current hypotheses, post unresolved issues, recompute evaluations, etc. The expert's knowledge of how to do analysis in the task is also represented in rule form. These strategy rules replace the normal executive program.

The situation-hypothesis is represented as a node-link graph, tree-like in that it has distinct "levels," each representing a degree of abstraction (or aggregation) that is natural to the expert in his understanding of the domain. A node represents an hypothesis; a link to that node represents support for that hypothesis (as in HEARSAY II, "support from above" or "support from below"). "Lower" levels are concerned with the specifics of the signal data. "Higher" levels represent symbolic abstractions.

3.4.4 Sketch of method

The situation-hypothesis is formed incrementally. As the situation unfolds over time, the triggering of rules modifies or discards existing hypotheses, adds new ones, or changes support values. The situation-hypothesis is a common workspace ("blackboard," in HEARSAY jargon) for all the rules.

In general, the incremental steps toward a more complete and refined situation-hypothesis can be viewed as a sequence of local generate-and-test activities. Some of the rules are plausible move generators, generating either nodes or links. Other rules are evaluators, testing and modifying node descriptions.

In typical operation, new data is submitted for processing (say, N time-units of new data). This initiates a flurry of rule-triggerings and consequently rule-actions (called "events"). Some events are direct consequences of the data; other events arise in a cascade-like fashion from the triggering of rules. Auxiliary symbolic data also cause events, usually affecting the higher levels of the hypothesis. As a consequence, support-from-above for the lower level processes is made available; and expectations of possible lower level events can be formed. Eventually all the relevant rules have their say and the system becomes quiescent, thereby triggering the input of new data to re-energize the inference activity.

The system uses the simplifying strategy of maintaining only one "best" situation-hypothesis at any moment, modifying it incrementally as required by the changing data. This approach is made feasible by several characteristics of the domain. First, there is the strong continuity over time of objects and their behaviors (specifically, they do not change radically over time, or behave radically differently over short periods). Second, a single problem (identity, location and velocity of a particular set of objects) persists over numerous data gathering periods. (Compare this to speech understanding in which each sentence is spoken just once, and each presents a new and different problem.) Finally, the system's hypothesis is typically "almost right," in part because it gets numerous opportunities to refine the solution (i.e. the numerous data gathering periods), and in part because the availability of many knowledge sources tends to over-determine the solution. As a result of all of these, the current best hypothesis changes only slowly with time, and hence keeping only the current best is a feasible approach.

Of interest are the time-based events. These rule-like expressions, created by certain rules, trigger upon the passage of specified amounts of time. They implement various "wait-and-see" strategies of analysis that are useful in the domain.

3.4.5 Results

In the test application, using signal data generated by a simulation program because real data was not available, the program achieved expert levels of performance over a span of test problems. Some problems were difficult because there was very little primary signal to support inference. Others were difficult because too much signal induced a plethora of alternatives with much ambiguity.

A modified SU/X design is currently being used as the basis for an application to the interpretation of x-ray crystallographic data, the CRYSALIS program mentioned later.

3.4.6 Discussion

The role of the auxiliary symbolic sources of data is of critical importance. They supply a symbolic model of the existing situation that is used to generate expectations of events to be observed in the data stream. This allows flow of inferences from higher levels of abstraction to lower. Such a process, so familiar to AI researchers, apparently is almost unrecognized among signal processing engineers. In the application task, the expectation-driven analysis is essential in controlling the combinatorial processing explosion at the lower levels,exactly the explosion that forces the traditional signal processing engineers to seek out the largest possible number-cruncher for their work.

The design of appropriate explanations for the user takes an interesting twist in SU/X. The situation-hypothesis unfolds piecemeal over time, but the "appropriate" explanation for the user is one that focuses on individual objects over time. Thus the appropriate explanation must be synthesized from a history of all the events that led up to the current hypothesis. Contrast this with the MYCIN-TEIRESIAS reporting of rule invocations in the construction of a reasoning chain.

Since its knowledge base and its auxiliary symbolic data give it a model-of-the-situation that strongly constrains interpretation of the primary data stream, SU/X is relatively unperturbed by errorful or missing data. These data conditions merely cause fluctuations in the credibility of individual hypotheses and/or the creation of the "wait-and-see" events. SU/X can be (but has not yet been) used to control sensors. Since its rules specify what types and values of evidence are necessary to establish support, and since it is constantly processing a complete hypothesis structure, it can request "critical readings" from the sensors. In general, this allows an efficient use of limited sensor bandwidth and data acquisition processing capability.

3.5 OTHER CASE STUDIES

Space does not allow more than just a brief sketch of other interesting projects that have been completed or are in progress.

3.5.1 AM: mathematical discovery

AM is a knowledge-based system that conjectures interesting concepts in elementary mathematics. It is a discoverer of interesting theorems to prove, not a theorem proving program. It was conceived and executed by D. Lenat for his Ph.D. thesis, and is reported by him in these proceedings ("An Overview of AM").

AM's knowledge is basically of two types: rules that suggest possibly interesting new concepts from previously conjectured concepts; and rules that evaluate the mathematical "interestingness" of a conjecture. These rules attempt to capture the expertise of the professional mathematician at the task of mathematical discovery. Though Lenat is not a professional mathematician, he was able successfully to serve as his own expert in the building of this program.

AM conducts a heuristic search through the space of concepts creatable from its rules. Its basic framework is generation-and-test. The generation is plausible move generation, as indicated by the rules for formation of new concepts. The test is the evaluation of "interestingness." Of particular note is the method of test-by-example that lends the flavor of scientific hypothesis testing to the enterprise of mathematical discovery.

Initialized with concepts of elementary set theory, it conjectured concepts in elementary number theory, such as "add," "multiply" (by four distinct paths!), "primes," the unique factorization theorem, and a concept similar to primes but previously not much studied called "maximally divisible numbers."

3.5.2 MOLGEN: planning experiments in molecular genetics

MOLGEN a collaboration with the Stanford Genetics Department, is work in progress. MOLGEN's task is to provide intelligent advice to a molecular geneticist on the planning of experiments involving the manipulation of DNA. The geneticist has various kinds of laboratory techniques available for changing DNA material (cuts, joins, insertions, deletions, and so on); techniques for determining the biological consequences of the changes; various instruments for measuring effects; various chemical methods for inducing, facilitating, or inhibiting changes; and many other tools.

MOLGEN will offer planning assistance in organizing and sequencing such tools to accomplish an experimental goal. In addition MOLGEN will check user-provided experiment plans for feasibility; and its knowledge base will be a repository for the rapidly expanding knowledge of this specialty, available by interrogation.

Current efforts to engineer a knowledge-base management system for MOLGEN are described by Martin et al in a paper in these proceedings. This subsystem uses and extends the techniques of the TEIRESIAS system discussed earlier.

In MOLGEN the problem of integration of many diverse sources of knowledge is central since the essence of the experiment planning process is the successful merging of biological, genetic, chemical, topological, and instrument knowledge. In MOLGEN the problem of representing processes is also brought into focus since the expert's knowledge of experimental strategies -- proto-plans -- must also be represented and put to use.

3.5.3 CRYSALIS: inferring protein structure from electron density maps

CRYSALIS, too, is work in progress. Its task is to hypothesize the structure of a protein from a map of electron density that is derived from x-ray crystallographic data. The map is three-dimensional, and the contour information is crude and highly ambiguous. Interpretation is guided and supported by auxiliary information, of which the amino acid sequence of the protein's backbone is the most important. Density map interpretation